When Users Control the Algorithms: Values Expressed in Practices on the Twitter Platform

JENNA BURRELL, School of Information, University of California, Berkeley, USA
ZOE KAHN, School of Information, University of California, Berkeley, USA
ANNE JONAS, School of Information, University of California, Berkeley, USA
DANIEL GRIFFIN, School of Information, University of California, Berkeley, USA

Recent interest in ethical AI has brought a slew of values, including fairness, into conversations about technology design. Research in the area of algorithmic fairness tends to be rooted in questions of distribution that can be subject to precise formalism and technical implementation. We seek to expand this conversation to include the experiences of people subject to algorithmic classification and decision-making. By examining tweets about the "Twitter algorithm" we consider the wide range of concerns and desires Twitter users express. We find a concern with fairness (narrowly construed) is present, particularly in the ways users complain that the platform enacts a political bias against conservatives. However, we find another important category of concern, evident in attempts to exert control over the algorithm. Twitter users who seek control do so for a variety of reasons, many well justified. We argue for the need for better and clearer definitions of what constitutes legitimate and illegitimate *control* over algorithmic processes and to consider support for users who wish to enact their own collective choices.

CCS Concepts: • **Human-centered computing** → *HCI theory, concepts and models*; *Empirical studies in collaborative and social computing*.

Additional Key Words and Phrases: algorithmic fairness, human autonomy, control, automation, gaming the algorithm, assembly, Twitter

ACM Reference Format:
Jenna Burrell, Zoe Kahn, Anne Jonas, and Daniel Griffin. 2019. When Users Control the Algorithms: Values Expressed in Practices on the Twitter Platform. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 138 (November 2019), 20 pages. https://doi.org/10.1145/33592407

1 INTRODUCTION

Concern over algorithmic fairness has been the focus of recent scholarship, particularly in work that scrutinizes how machine learning algorithms are applied in high stakes domains such as criminal justice. These discussions often treat fairness as rooted in questions of distribution that are receptive to precise formalism and technical implementation [13, 15, 25, 34]. While this is perhaps closer to ways of thinking about fairness within some fields, for example economics, it is more distant from a broader range of legal and philosophical views aligned with questions of personhood and justice. These definitional debates about algorithmic fairness have also tended to omit the voices of users/stakeholders, with some key exceptions [17, 35, 51].

Authors' addresses: Jenna Burrell, jburrell@berkeley.edu, School of Information, University of California, Berkeley, USA; Zoe Kahn, zkahn@berkeley.edu, School of Information, University of California, Berkeley, USA; Anne Jonas, annejonas@berkeley.edu, School of Information, University of California, Berkeley, USA; Daniel Griffin, daniel.griffin@berkeley.edu, School of Information, University of California, Berkeley, USA.

2573-0142/2019/11-ART138
https://doi.org/10.1145/33592407

We draw from public social media data, specifically tweets that mention "Twitter's algorithm" (and variants of that phrase) as a starting point to understand the concerns users raise and practices they undertake when they explicitly reference the algorithm. We found that users expressed fairness concerns when they perceived that Twitter was granting its core resource – attention – in a way that advantaged some groups over others. However, we surfaced another set of concerns that were largely distinct from fairness. These concerns related instead to users' sense of being controlled by or seeking control over Twitter's algorithms. When such issues have been raised within the community concerned with the ethics of algorithmic systems, they are generally labeled as matters of human autonomy or as dignitary considerations [31, 33, 43]. However, this area of concern has received relatively less emphasis in emerging scholarship. In fact, discussions of algorithmic fairness seem to often tack in the opposite direction, toward developing tools that better mediate the fair allocation of resources based on expert definitions and new forms of technical validation.

Twitter is one of many online platforms which deploy algorithmic processes to manage, rank, filter, and in other ways organize core elements of what users see. Scholars have referred to these as "curation algorithms" [17] that have the power to "grant visibility and certify meanings" [21]. This curation is carried out largely as a background process and users are generally expected to influence these algorithms passively. Their behavioral traces (such as 'liking' or 'retweeting,' clicking on links, viewing pages or profiles) serve as inputs shaping the dynamics of algorithmic classification. When users attempt to interact with these algorithms directly and intentionally, bringing to bear an awareness of how they work, they are exerting a kind of control that is not explicitly granted by the platform.

Our collection of tweets that reference "Twitter's algorithm" served as a valuable index to a wide range of sub-communities and platform practices. In their efforts to control algorithmic processes, we found that users treat Twitter as an important site for assembly. Users developed rich strategies and used a host of techniques, many that are explicitly algorithmically-aware, to define and manage their community and its boundaries within Twitter, to get their message out and find a wider audience, to make-do where the platform fails them, and to resist the platform's intrusions. Within this are forms of control over the algorithm that are deeply troubling or that are wielded unjustly against other platform users. Yet, in many cases, and as our data show, control strategies can also have a valuable purpose, are often collectively developed and enacted, and have considerable social utility. Alongside fairness we argue that control (and by extension user autonomy) are equally important values to account for in the design of algorithmic processes. This is not just because control strategies can have social utility, but because control is itself an important dignitary concern.

2 BACKGROUND, RELATED WORK

The recent surge in interest over ethical AI has renewed and intensified consideration of a slew of values, including fairness, in technology design (i.e. [5, 39]). This is a continuation of work set in motion over two decades ago [19, 42]. Recent work has sought to define and implement fairness algorithmically, but the quest for a fair algorithm has been impeded by competing definitions. The concern with individual fairness, for example, suggests that similar individuals, should be classified similarly [32]. By contrast, group fairness aspires to statistical parity in outcomes for different demographic groups. Fairness can be considered a matter of disparate treatment versus disparate impact [3, 37]. The former means that individuals should be subject to the exact same algorithmic process (the treatment) whereas the latter suggests fairness is determined by the end outcome (the impact). Scholarship has shown through formal mathematical proofs that all definitions of fairness cannot be satisfied by one optimally fair algorithm [34]. While fairness discussions have drawn the attention of computer science theorists and economists in particular, questions of algorithmic

control, because they implicate user-facing design questions, call out to other technical specialties including Human-Computer Interaction (HCI).

While 'control' has not been a central focus in the recent revival of ethics concerns that pertain to algorithmic systems, there is a long history of work in HCI that has examined questions of automation and of designing human controls into AI systems [27–29]. Such work often tackles this from a cognitive perspective, arguing for opportunities to enhance (rather than replace) human intelligence through productive human-machine partnerships. However, we turn attention instead to arguments from legal scholarship that forward user/stakeholder control over algorithmic processes, including through collective actions, as linked to the values of accountability and human autonomy.

We draw from Beniger's [4] definition of control as "purposive influence toward a predetermined goal." In this paper we seek to explore how and why those who are subject to algorithmic classification attempt to exert control over such systems. That there is "general value" in "empowering people to navigate the algorithms that affect their lives" seems largely self-evident [43]. It is furthermore already encoded in certain legal domains. For example, in the U.S., the Fair Credit Reporting Act of 1970 set regulations around credit scoring and reporting. Specific reasons for a denied loan must be provided to consumers, in part, to "educate the consumer about how such a result could be changed in the future" [43]. Following the Equal Credit Opportunity Act of 1974, membership in a protected class can not be used for a credit denial. Protected classes cover aspects of individual identity such as race or ethnicity, gender identity, age, and religion, that an individual may be unable (or should not be asked) to change. Kaminski [33] also locates control (or more broadly autonomy) within discussions of algorithmic decision-making as a dignitary concern and links secrecy about such mechanisms to this lack of control. In European political culture, questions of control are more fundamentally tied to the fact of automation as reflected in the recent General Data Protection Regulation (GDPR). Where there is no human capacity to intervene in algorithmic decision-making (whether by the person subject to the decision or by a human intermediary) this is understood as a fundamental threat to human dignity [31].

Conversely, another line of work defines how control over algorithmic processes by those subject to them may be undesirable in some instances. Illegitimate control is often referred to as *gaming* the algorithm. Work to define gaming is fairly extensive in computer science including a growing literature on adversarial machine learning[1]. Hardt et al. [24] define gaming as the exposure of an algorithmic classifier to "public scrutiny" leading to strategic behavior that produces "unintended advantage." This degrades the accuracy of the classifier over time and is a major concern when such a system is used to determine the allocation of scarce resources. In legal scholarship, 'gaming' has been defined in a slightly different way. Bambauer and Zarsky [2] note that assessment or classification algorithms rely on input proxies that do not correspond exactly with the thing they represent. For example, SAT scores are a proxy for college readiness. This gap provides an opportunity for gaming which occurs when the proxy is altered without changing what the proxy represents. Importantly, attempts to exert control over algorithms are not automatically gaming by either of these definitions. For example, changing the source behavior a proxy represents, i.e. paying off debt to become more creditworthy, would be a way of exerting control over a credit scoring algorithm without gaming it.

There are also robust arguments made in defense of gaming. Bambauer and Zarsky [2] state, "people who engage in gaming can represent the full range of ethical motives, from parasitic, to benign, to downright noble." They note that it produces, "benefits that stem from fundamental values of dignity and autonomy." Gaming as unintended use "can spur innovations," help to "find

[1] [44] offers an applied example in social media

novel uses for existing platforms" and provide "opportunities to play as opposed to being constantly 'systemitized'." Other defenses of gaming argue that practices specifically of user-driven "obfuscation" are a form of principled resistance and self-protection against systems of pervasive data gathering and surveillance [7]. While enhancing control may come at the cost of classifier accuracy, [2] suggest that it may be justified nonetheless when weighing other values.

Some scholars note that control, particularly where defined as 'gaming,' intersects with concerns over algorithmic fairness as well. Gaming requires resources of knowledge and time and may consequently undermine the equitable distribution of a resource. In [2] the authors note that "some populations will be less willing or able to engage in gaming, and therefore both gaming and countermoves can have disparate effects on different subgroups." In the literature so far cited we can see a number of ways in which concerns over transparency, fairness, and control are overlapping and interlinked.

Past studies of social media platforms have likewise sought a reconsideration of user control over algorithmic processes that have been discredited as *gaming* [8]. Such studies recognize that attention is a primary resource users seek to manipulate strategically. Cotter [14] defines the practices of Instagram "influencers" seeking to build a following and garner attention as "playing the visibility game." However, work on how platform users exploit strategies to maximize attention offers insight into only one dimension of gaming behavior [6, 14, 21]. Other studies have conversely found strategies of algorithmic manipulation that attempt to reduce visibility, to draw *less* attention rather than more. Along these lines, Van der Nagel [47] considers "forced connections" where platform algorithms connect "otherwise disparate data points on digitally networked platforms to the subject's detriment." Users respond by pursuing "efforts to wrest back control." In this article we bring together a variety of phenomena, including some documented in past studies of social media, to show how strategies of algorithmic control are a common thread.

Twitter and other social media platforms play a key role as arbiters of their computational resources. Consequently, the perspective of such firms will likely be especially focused on gaming and on distinguishing legitimate and illegitimate forms of control. We may also treat the platform's business interests as a factor in the way user control is shaped, tolerated, resisted, or facilitated. For social media platforms that are ad-supported (like Twitter) these interests are likely to center on ensuring consistent and accurate personal data collection, facilitating the monetization of that data, ensuring the unimpeded flow of ad revenue, and maximizing use to boost the value of the platform for advertisers. Formalized technical definitions of gaming, such as [24] likewise focus on harm to a particular tool (a classifier algorithm). While ways of defining gaming offer an important commentary on control and caveats about its value, this paper centers on the perspective of users on what control facilitates for them and for other groups who are impacted by algorithmic processes.

3 METHODS

3.1 Dataset Construction and Analysis Procedure

As an initial step in our study, we used the Twitter API to collect tweets from September 28, 2018 to January 18, 2019 [2]. Our primary dataset contained tweets that use the phrase "Twitter algorithm" and its variants. These variants include differences in capitalization ('Twitter' vs. 'twitter'), possessive and non-possessive (Twitter vs. Twitter's), and singular and plural ('algorithm' and 'algorithms'). We only included tweets in our dataset where these two terms were in immediate sequential order. In total, this dataset is comprised of 8875 tweets (4280 original, non-reply tweets and 4595 tweets

[2] we polled Twitter's standard API with our search term every 30 seconds configured for 'recent' rather than 'top' (high relevance) results. We believe the relatively low frequency of our search term means that our dataset is close to comprehensive in this period, see [10]

Table 1. total tweets in dataset

	"twitter algorithms"
non-reply tweets	4280
reply tweets	4595
total tweets	8875
unique user accounts	8041

in reply to other users) from 8041 distinct user accounts. As the term "algorithm" has begun to circulate in more mainstream use appearing in news coverage, opinion pieces, mass market books and television, and other spaces of public debate, studies such as ours have become more feasible. This approach is starting to be leveraged by other scholars as well [8]. While we initially explored other search terms to surface user commentary about the Twitter platform including the phrases, "who to follow[3]," "shadowbanning[4]," and tweets directed at Twitter's head of trust and safety, Del Harvey (@delbius)[5], the "Twitter algorithm" dataset gave us both breadth (tweets that cut across a wide variety of topics and concerns) and high relevance given our research interests.

Our initial process of analysis involved a two-part coding procedure primarily intended to help us organize and work through the data systematically. We drew from elements of grounded theory (by starting with an open-ended inductive coding process) as well as content analysis (in our use of multiple coders and inter-coder reliability calculations). Our first step was to iteratively develop a code book by inductively coding around one hundred tweets to discover major recurring themes in the tweets. We used a strategy from grounded theory applying gerunds (action terms) to surface process, what the tweets themselves seemed to be doing [12]. As certain codes emerged with relative frequency, we narrowed our focus arriving at a streamlined codebook with 10 codes grouped within four broad categories (explanations, complaints, relationship management, and controlling/gaming the algorithm). In the second step, we engaged in a more comprehensive process of attempting to code a sample of 911 tweets using this codebook. Researchers met twice to discuss and reconcile how they had coded the same tweets using the 10 codes[6].

The sample included the first 10 out of every 100 tweets in the "Twitter algorithms" dataset. Tweets were listed in chronological order, so the sampling technique allowed us to move through the data in such a way that particular events in the time period did not overwhelm our analysis[7]. About 11% of the tweets were no longer public and were subsequently excluded from our dataset. This included tweets from suspended or protected accounts and tweets that had been deleted. Some tweets we deemed uninterpretable. This percentage ranged from 5-11% depending on the coder. Uninterpretable tweets were ones that contained too little information, replied to no-longer available tweets, used personal names of people we did not know or cultural references we could not track down, or were generally nonsensical. Our final coded dataset included 718 tweets.

[3] 'who to follow' recommendations are prominent in the Twitter user interface and are produced by a recommendation algorithm

[4] shadow banning was a popular theory particularly among political conservatives about how Twitter algorithmically singled them out denying the platform's attentional resources but in a way that was intentionally hard to detect

[5] tweets to Del Harvey were not specifically about the platform's algorithms, but were comparable for their focused commentary on the platform and its operation

[6] Our reported percentages use the assessments of the primary coders (the first two authors) who each coded half of the dataset sample

[7] which included the U.S. midterm election, President Trump complaining about mass loss of Twitter followers (after an earlier shift by Twitter to remove locked accounts from follower numbers), and the discovery that a mail bomb suspect had previously threatened public figures on Twitter

For each tweet in our sample, we read the full text of the tweet, viewed the tweet on the twitter platform to consider the context of the message including retweeted content, its embedding in a tweet thread, the original tweet the user was replying to, and images, links, screenshots or other non-textual supplementary materials. For each coded tweet we gave a true or false value for each of the ten codes depending on whether that theme appeared to be present or not in the tweet. Each tweet was coded independently by two coders. Kappa coefficients measuring inter-coder reliability above chance agreement ranged from fair to good (50% to 88%) [41]. A full description of the ten codes used by the coders along with the frequency of their occurrence and Kappa coefficient scores are in the appendix.

3.2 What Claims Can (and Cannot) Be Made From This Dataset

Our approach draws generally from a non-positivist approach to interpretation. Thus our findings below focus not on claims that rest on the proportionality of codes across themes or other forms of quantification. Instead we focus on the diversity *within* the themes and explore this through example tweets. Looking at the tweets within code categories allows us to compare the way Twitter users speak about algorithmic processes with scholarship on algorithms and machine learning ethics that roughly equate to those code categories. We find that examples of users seizing control of algorithmic processes (or complaining about the lack of control) are in ready evidence. Thus, *control*, the focus of our analysis below, is not a niche theme or an imagined one, but has a demonstrable empirical basis and some measure of inter-subjective validity. In presenting the primary data below (in the form of the full text of tweets) we invite readers to assess for themselves our interpretation of control themes.

The tweets in our dataset generally referenced the recent experiences of users on the platform and had a certain quality of immediacy. Thus the frequency of codes indicates the events on the platform that draw the attention of users, specifically events that provoke reflection on the platforms 'algorithms.' The tweets in our dataset tended to focus (unsurprisingly) on the most visible parts of the platform. For example, the curation algorithm that determines what tweets a user will and won't see in their timeline was discussed in 38% of coded tweets in our dataset, far more than any other nameable algorithm[8]. Importantly, we do not consider code frequencies to offer precise measurement of user's attitudes or priorities (i.e. whether algorithmic fairness or control is more important to them).

3.3 Limitations and Future Directions

We recognize that our dataset is not representative of any group of individuals and certainly not representative of Twitter users as a whole [26, 45]. We expect, based on their interest in platform curation and content moderation practices and familiarity with the still somewhat technical term "algorithm," that our dataset represents the views of a somewhat more dedicated and technically knowledgeable base of Twitter users.

The 'population' we are sampling from is the texts (in this case, tweets) not the users tweeting those texts. This alternative view of the research 'population' is established in standard definitions of content analysis [41]. An investigation into *who* tweets about the Twitter algorithms could be a productive line of research inquiry for future study following a similar method pursued in [8]. Social network analysis of follower/following links between accounts in our dataset or accounts they share in common would also be an illuminating sequel to this study.

[8]while we coded each tweet for the specific algorithm it seemed to reference the inter-coder reliability for this was very low. Furthermore in many tweets the algorithm referenced was ambiguous. Twitter also does not publish a list of content curation or content moderation algorithms it employs making it difficult for users and researchers alike to label specific algorithms.

On the matter of ethics, following insights from [18], for the subset of tweets we published in their entirety in this paper, we reached out to their authors to ask whether they wanted their tweets anonymized or not, or whether they wanted them excluded entirely from our paper. Unless authors told us otherwise we presumed a preference for anonymity and altered tweet language in ways designed to evade searchability. As a way to protect Twitter users from further and unnecessary privacy intrusion, we did not seek out user profile information, previous tweets by these accounts, their presence on other platforms, or details about the offline lives of users. This additional user information we determined was not necessary to answer our research questions that had to do with how control is evidenced in tweets about the "Twitter algorithm."

4 FINDINGS

In our dataset, an awareness of algorithmic processes among Twitter users is a given. Prior research has examined whether or not users are aware of the algorithmic curation of content online finding that many users are *not* aware of the algorithmic processes that shape their online experience [16, 48]. We were instead interested in the moment where knowledgeable users turn their attention to the algorithms they perceive are influencing their platform experience and in the commentary they provided in these moments.

4.1 Invoking Twitter 'algorithms': meta-categories

Under what circumstances do people explicitly reference Twitter's algorithms? We arrived at four meta-categories in our inductive coding process and within these categories we developed 10 distinct codes. The four meta-categories are as follows:

(1) Some tweets **sought or offered explanations** about Twitter algorithms. These included tweets that puzzled over the algorithm seeking an explanation from other users. Other tweets in this category provided such an explanation, sometimes in direct response to a question posed in another tweet. Many of these tweets had no clearly stated end goal for such explanations (i.e. an action that would be taken based on better understanding the algorithm). Others were highly directive with explanations that seemed intended to alter the behavior of other users.
(2) A considerable proportion of tweets were **complaints** about Twitter algorithms. We coded 51.9% of tweets as expressing negative sentiments such as problems with or a dislike of Twitter algorithms while 40.5% were neutral and a mere 7.6% expressed positive sentiment. Bad experiences on the platform seemed to motivate a disproportionate amount of tweeting about Twitter algorithms.
(3) A third set of tweets we categorized as **relationship management**. Many of these tweets attempted to shift blame to the algorithm for some damage to the bond or ties between users on the platform. This was undertaken by users to explain why they had seemingly ignored another user's tweets, to reassure fans lamenting a lack of attention from a celebrity, and by users trying to seem less 'creepy' when replying to strangers. The tweets in this category highlighted the work users undertook to manage and maintain the interpersonal connections they develop on the platform. It helped us to see the value of Twitter as a place of assembly. This proved important as we later came to see how it animated much of the work done by users to exert control over Twitter's algorithms. An example tweet that reflects this sensitivity to how the platform algorithms can disrupt personal connections:

> "When I realize I haven't seen a Twitter friend tweet in a while. But, I realize that might be because Twitter's algorithm hasn't been surfacing their tweets. So, I do an actual

search for their account and sure enough they've been tweeting good tweets I haven't been supporting."

(4) Finally, a fourth category of tweets were attempts to **exert control over** Twitter platform algorithms. This theme was present in 17% of coded tweets. In some cases, strategies of control were embedded in the tweet itself. Other tweets offered explanations of how the algorithm works (as in meta-category one above) but with clear instructions for how to translate that into actions that would produce preferable results. This final meta-category we elaborate upon later in the paper.

While our main focus in this paper is on how 'control' surfaced as a theme in our dataset, we wish to elaborate further on tweets labeled with codes from the 'complaints' and 'relationship management' meta-categories for several reasons. First, the tweets in the complaints meta-category map roughly to existing issues raised in scholarship on algorithmic systems and machine learning. There is some correspondence between the problem framing in scholarship and how users frame their concerns, but also some points of divergence. Second, both of these meta-categories illuminate practices on the platform that are relevant to the control theme. Finally, the 'relationship management' meta-category foregrounds how Twitter functions as a sociable space which we will return to later in the article in understanding social group formation on the platform.

In complaints about the algorithm, we are able to better understand how users experience the platform and how they frame its key shortcomings in relation to their first-hand experiences. Some complaints were about accuracy and errors. This is a major emphasis in the machine learning literature where accuracy (particularly predictive accuracy) is valued as the primary assessment criterion. On Twitter this complaint was grounded in dissatisfaction with what content (tweets, ads, who to follow suggestions) was served to them in their timeline and other areas of the user interface. Generally when users complained about accuracy in this way, they reflected a sense that Twitter content was insufficiently personalized. For example, one recurrent complaint about being served advertisements for luxury goods (i.e. Rolex watches) came from users who perhaps either reject conspicuous consumption or have limited financial means. In other words, they expected the platform to know them better in serving ads and other algorithmically-determined content.

Unfairness or bias also registered for some users as a problem. We developed a distinct code within the "complaints" meta-category to identify such tweets. In our dataset, such complaints predominantly came from the Twitter accounts of political conservatives complaining about the platform's political bias against them[9]. In these tweets we see that "fairness" does indeed register as a concern for Twitter users, but in a way that is mostly focused on perceptions that Twitter executives or the company as a whole has a political agenda. However, there were also complaints by Twitter users living outside of the US about regional bias, accusing Twitter algorithms of directing disproportionate attention to US-based news and events.

This category of complaint maps most closely to the particular literature on fairness that centers on issues of resource distribution. We see in these complaints a recognition that Twitter algorithms are tasked with distributing its scarce but valuable resource – attention. Users perceive that it does so unevenly, in a way that privileges members of one group over another. We are skeptical, however, that an algorithm mathematically validated as fair by some formal definition could resolve this particular complaint. Furthermore, fairness, particularly in the sense of distribution, may not be something individual users are well positioned to perceive because it requires access to information about overall patterns across the system. While potentially a problem, it is possibly less likely to

[9]We found claims about Twitter's bias against political conservatives in the majority (58%) of tweets we coded as 'bias complaints'.

register in the experiences of users or be reported or complained about directly. In this vein, some users' 'bias' complaints fall closer to concerns around censorship than fairness.

Issues related to user control and autonomy emerged across several code categories. We developed a particular code for control within the "complaints" meta-category. In tweets coded as control complaints, users lamented the way Twitter algorithms controlled their behavior on the platform. This sometimes included accusations of censorship such as disallowed language that triggered the platform's content moderation processes potentially leading to account suspension. Generally these users perceived 'algorithms' as part of some kind of automated system that detected sensitive keywords. They also complained about the way platform algorithms facilitated users to exert control over others. They complained that reporting functions were wielded against them or others for illegitimate purposes:

> "Certain people are flagging tweets containing violent words like "murder" and "kill." Twitter's algorithm, of course, can't distinguish roleplay from real-life threats. Someone's exploiting this. They want to suspend accounts they have a problem with. So fucked up."

In a lecture, Sara Ahmed [1] makes the point that the act of complaint itself can be a way for people to record their grievances and build solidarity in the face of limited recognition by those with organizational power. We will also show that many of the strategies of exerting control that our dataset surfaced reflected power asymmetries that disprivileged users in relation to the platform.

Issues related to control appeared again in tweets we coded under the "relationship management" meta-category. This category showed the work done by users to repair relationships in the aftermath of algorithmic processes. We describe this as control *around* the algorithm. We found users both pro-actively preventing relationship damage and reactively repairing it by shifting blame from humans to the algorithm, and also occasionally shifting blame from the algorithm back to humans. Some of this stemmed from the challenge of managing a multitude of relationships and identities within a platform that treated engagement and attention in singular ways without regard for Twitter user's possibly disparate audiences, a problem elsewhere described as "context collapse" [38]. For example, this tweet expresses an apology (possibly tongue in cheek) for exposing Twitter followers to content (gay porn) that might not be of general interest:

> "I deeply apologize to anybody who's seen gay porn as a due to me 'liking' it. Twitter algorithms are hard at work once again."

Another tweet reflects an awareness of the reciprocity expectations among some users while highlighting a constraint imposed by the platform (an automated cap on follows for some accounts) that prevented them from participating:

> "New followers - I will follow back as soon as I can. The twitter algorithm won't let me follow more people right now.
> *waves fist at twitter algorithm*"

Some of these relationship management tweets reflect the way users work to control an, "understanding of the symbiotic agency in human and algorithmic communication" by reminding others on the platform how their behavior is read and shifted by the algorithms [40]. Most of the control-related tweets in our dataset (17%), however, fell under the fourth meta-category. This category encompassed strategies to enact control over algorithmic processes which we will elaborate next.

4.2 Exerting Control Over the Algorithm

By examining the tweets we coded under the "exerting control" meta-category we can better understand the value and utility of control according to platform users. These tweets show users

actively attempting to influence Twitter platform algorithms to benefit themselves, their Twitter account, a Twitter-based subcommunity they belonged to, or a particular topic or cause. These efforts went beyond explanations (simply trying to understand the algorithm) or complaints about wanting or lacking control over the algorithms. In some cases, attempts at control were embedded in the content of the tweet itself, through careful word selection, hashtag use, word omission or other coded language. We found attempts by users to make their accounts, tweets, identity, or topics more visible to the algorithm, to attract followers, or to get their message out, by becoming more "algorithmically recognizable" [21]. However we also found many attempts to accomplish the opposite - to be algorithmically unrecognizable, through "obfuscation" [7], dodging [50], or other tactics of resistance [47]. This was sometimes done to minimize scrutiny from the platform (and its content moderation processes) or to avoid attention from certain individual users or groups on the platform.

On Twitter, to exert control of any sort, users had limited resources at their disposal. They had the ability to arrange freeform text in tweets (limited to 280 characters), to form tweet threads or post images of text that overcame those character limits, to use hashtags, and employ the platform-provided user interface features (liking or retweeting tweets, following and unfollowing other accounts, reporting tweets or accounts for site policy violations). They were also sometimes more creative, pushing beyond what these formal UI features enabled. Importantly, users had networks of other users (especially their followers) as a resource to draw from. They used tweets to call out to other users and to coordinate efforts.

We have organized the control strategies undertaken by users that surfaced in our data into three subsections reflecting distinct practices. These strategies are: (1) personalization against platform defaults, (2) working around platform errors and limitations, and (3) virtual assembly work. In examining these distinct strategies we are able to define more clearly the specific form taken by user control efforts on Twitter. Many examples on the surface level appear innocuous or justifiable. Digging deeper into existing scholarship about how online spaces are evolving as a public good, we find that users' control strategies are also intertwined with certain values and ideals that have been argued for in legal scholarship. We suggest that an overfocus on 'gaming' and a concern by firms like Twitter with adversarial dynamics may undermine productive and meaningful forms of user control.

4.3 Personalization Against Platform Defaults

Individual efforts to "exert control" often included taking action to influence what content was displayed to them alone, to improve content personalization (e.g., by strategically following, unfollowing or muting accounts; by systematically 'liking' or 'retweeting' particular kinds of content thought to produce a desired result). These efforts don't monopolize platform resources and they don't degrade the experience for other users and so they don't fit any definition of gaming we have so far considered. In some of these tweets users attempted to disrupt the 'filter bubble,' a perceived process of algorithmic curation where content selected for a user confirms and reinforces their worldview and ideology rather than presenting challenging or conflicting viewpoints. We've noted that tweets coded under the 'complaints' code often highlighted the insufficient personalization of Twitter content. This suggests that users expect content to closely reflect their sense of themselves. However this alternative attempt to 'depersonalize' the algorithm shows diverging views among users.

Other efforts at personalization do sometimes show an attempt to mildly undermine the platform's business interests. For example, some users sought ways to remove all ads, though its questionable whether a strategy that simply uses a Twitter provided feature (muting ads) would have much effect:

> "can feel twitter's algorithm growing more and more panicked as I mute more accounts that give me ads."

Other efforts sought to counter the platform logic of knowing users better in order to provide the tweets in their timeline that users "are likely to care about most."[10]. Along these lines, a number of tweets in our dataset reflected a desire by users to interfere with the construction of their "data double" [22] disrupting intrusive data surveillance practices (see [7]). In the following tweet, a user described their Twitter settings page which suggested a chaotic list of 'interests':

> "Decided to investigate what twitter algorithms think of me. I haven't it checked before on this account. Damn, whatever I'm doing I'm doing right as it's all over the place."

Such tweets illustrate resistance to a widespread algorithmic logic of profiling users in ever greater detail, a practice undertaken by many social media platforms. As van der Nagel argues citing Willson [47, 49], "the goal of any platform is easy to understand, as it is rooted in neoliberal systems of profit. Increasing the number of users and the time they actively spend on the platform is paramount, as this leads to data creation, which facilitates more comprehensive personal profiles that platforms can sell to advertisers, so they can target their products to the individuals and groups most likely to buy them." This direct attempt to undermine the accuracy of their data profile is therefore, "lightweight 'gaming' as resistance, a way of preserving personal autonomy [2]."

Among Twitter users, as one might expect on any massively scaled global platform, there are diverse views and opinions including diverging viewpoints about whether receiving personalized content by contributing personal data is a fair exchange. The Twitter platform provides very few explicit levers for configuration. As a result, users turn to the curation algorithms to try to manipulate background processes to realize envisioned ends that diverge from what the platform prescribes.

4.4 Working Around Platform Errors and Limitations

According to many tweets in our dataset, Twitter platform algorithms have a noted problem with reading sentiment and context. This problem implicates content curation as well as content moderation processes. For example, a previously quoted tweet referred to how "Twitter's algorithm...can't distinguish roleplay from real-life threats" when flagging accounts for suspension. While content curation is about delivering to users more of what the platform determines is engaging (i.e. keeps them on the platform), content moderation is, conversely, about suppressing or removing content that goes against site policies on posting violent, offensive, or sensitive material. In one tweet in our dataset, a control strategy to overcome such limitations proposed the use of a substitute word in place of one thought to be algorithm-triggering:

> "So Pro tip. Since Twitter's algorithm is being hyper aggressive with suspending accounts that use the word kill. Sarcastically I suggest we switch to the word kale. I swear to God Janet if you don't shut up I will kale. Kale me now
> Honestly in my mind it's more terrifying"

The platform algorithms' insensitivity to context also seemed to impact content curation. This surfaced in tweets that referenced political issues. Users sometimes replied to or retweeted tweets from the accounts of political figures in order to condemn or criticize their statements. Some complained about account recommendations that subsequently suggested those same accounts in their "who to follow" box on the main screen.

[10] https://help.twitter.com/en/using-twitter/twitter-timeline

Sentiment analysis as a technique remains rudimentary. Automated systems that can decode the nuance within the universe of human creative expression is a long, long way off. It is also not totally clear whether Twitter is simply unable to read the valence of political engagement or whether its algorithmic processes are designed to be indifferent to this. After all, engagement is of potential value to Twitter by keeping users on the platform irregardless of whether that engagement is a lively debate, escalating interpersonal hostility, or coordinated mob harrassment.

Recognizing that 'engagement' was the coin of the realm on Twitter and that negative attention to an account or a tweet still has a way of amplifying the original, disputed message, some tweets in our sample attempted to deny this kind of attention through their individual practices and by coordinating with others. For example there were 23 tweets by one user alone in our dataset, each a slight variation on this text:

> "@tedlieu @realDonaldTrump @POTUS Love your tweet but ... Retweeting Trump amplifies his message by making his tweet rise higher in the Twitter algorithm (more people will see it). If you must reference a Trump tweet please take a screen shot of his tweet and post it that way. Thank you @lolgop for this idea"

Along the same lines, this example from our dataset was a reply to a British television journalist:

> "@adamboultonSKY Cheers Adam, it's the advantages they get from the Twitter algorithm that most concerns us in the disinformation combat community - hence the advice to #UseScreenshots instead ??"

A search on the hashtag #UseScreenshots provides many examples of users admonishing other users to deny attention to adversaries and to do so by acting with an awareness of the ways that Twitter algorithms measure and amplify engagement. The workaround was to use screenshots (of tweets, for example) that critics want to discuss. This offered both a way to secure permanence (as widely criticized tweets often are deleted by users) but also to deny further attention, evade text searchability, and avoid attracting supporters of the original message. By developing the #UseScreenshots hashtag, users attempted to propagate and amplify this as a kind of best practice.

Prior research has similarly examined the mismatch between platform metrics or mechanisms and the needs and identities of users. Social media platforms have historically structured the identities of their users in inflexible ways. One example is Facebook's real names policy which created difficulties especially for trans people, abuse survivors, and Native Americans. This was widely contested and eventually lead to platform design changes which might be considered a form of delayed and indirect control through user feedback [23]. In a study of ride allocation algorithms on Uber and Lyft, Lee [36] found similar approaches of strategic control by drivers used to cope with overly simplistic mechanisms for driver-rider matching.

In the present study, we found that some Twitter users attempted to use the platform as a way to contribute support for causes and to help amplify important messages, but found it was easy to be undermined by the algorithms. Replying with criticism or condemnation to the account of a political figure could, nonetheless, help to reward that account's message by adding to the cumulative attention it received. This may be a shortcoming of the Twitter platform as currently designed perhaps related to the nascent state of automated sentiment analysis or to a business logic that rewards engagement unidimensionally, or simply the absence of ways to explicitly signal disapproval such as an 'unlike' button.

4.5 Virtual Assembly Work

This section turns attention to the question of user control in relation to the specific role that Twitter (along with other major social media platforms) is coming to play as a public commons.

We show how control, while possible to defend as a value in its own right, may facilitate practices that reinforce other values as well.

In the previous section our examples alluded to how the 'control' strategies of many users focused on manipulating Twitter's core resource – 'attention.' Attempts by users to maximize their share of attention on social media platforms, including through attempts to consciously control algorithmic processes, is a well-studied phenomenon. The Search Engine Optimization (SEO) industry is devoted to algorithmic manipulation of search engine rankings to maximize attention. Gillespie describes one campaign pursued by users to perform for the Google search engine in ways that would be "algorithmically recognizable" [21]. A study of instagram "influencers" [14], YouTube "Reply girls" [9], and of YouTube beauty vloggers [6] all examine strategies of knowing the algorithms as part of seeking the largest share possible of attention on the platform. Attention-seeking on some platforms is motivated, at least in part, by monetary rewards. For example, YouTube gives users who upload popular content a proportional cut of their ad sales.

On Twitter, users do not receive a cut of ad revenue or other direct rewards for having a large Twitter following[11]. The particular pattern of attention-seeking we observed was carried out by coordinated groups rather than individuals. Efforts along these lines included campaigns in which groups of users sought to boost attention for a shared 'object' of interest. We witnessed this in tweets that sought attention for musicians or music groups, dance crews, or other celebrities aimed ideally at placing the group name or a hashtag in the coveted Twitter Trends box on the platform's home page. Such campaigns may be timed around the release of a new music album or other event. One example of this surfaced in our dataset when an attempt by fans of a Korean boy band to get the music group on Twitter trends went awry. A series of tweets referenced "algorithms" and proposed a change in strategy (a new hashtag) to restart the effort:

> "The hashtag was changed to #XIUMIN24x7 because of errors in Twitter algorithms in #TEMPO_XIUMIN"

> "@WWEXOL @weareoneEXO - What's the issue with the Twitter algorithm? - Maybe if we don't use the underscore. - Can we decide now what to trend instead? - Maybe combining member names with song titles? - Idk. #Xiumin24x7"

Campaigns were distinctive in the way they mobilized users around a cause with a short timespan and a focused goal. Social issues sharpened by news events are also time-linked and our firsthand experience as Twitter users also suggests that campaigns may be undertaken around social causes and political issues, although we did not find specific examples of this in our particular dataset. Users who coordinated their efforts in this way recognized that algorithmic processes on Twitter likely picked up on aggregate trends across multiple accounts. Influencing other users was another way, potentially a powerful one, to shape algorithmic effects.

Algorithmic control strategies are a resource used in building voluntary social groups on Twitter. Within the vast, formless space of the Twittersphere where accounts and tweets are public by default and where explicit user configurations are limited, control is sometimes about giving form to and forming boundaries around these groups. We found recurrent practices in our dataset of *avoiding* attention from certain groups, particularly given the threat of harassment on the platform. One example is the practice of "subtweeting" where users carefully left out an obvious account mention, typically that of a very public or well-known person with a large following. Subtweeting

[11] indirect monetary rewards certainly might motivate attention-seeking strategies on the Twitter platform as on other social media. The use of social media platforms has become core to the marketing strategies of professionals and small businesses.

and the use of coded language are intended to disrupt Twitter's system of automatic notifications and to keep a conversation more localized to their own account followers. For example:

> "Since twitter algorithms and progressives attribute hate where there is none, we may need to develop code words for any criticism of non-straight attraction. The term 'Idiosexual' might hold the thought police at bay for a bit."

Tufekci refers to this as "behavior aimed at algorithmic invisibility" [45] and van der Nagel [47] identified this as "voldemorting" where a key name or term is left un-named to avoid summoning adversaries (in just the way Lord Voldemort's name was unspoken in the Harry Potter series). She notes strategies around this related to the "Gamergate" controversy where the hashtag #GamerGate summoned coordinated mob harassment against women and other allies speaking out against misogyny in video game culture. Instead substitute terms like "Goobergate, Gatergate, GG or G@merg&te" were used.

While our literature review centered on how control was discussed and how it was defended in general terms or delegitimized as 'gaming,' our findings answer a more specific question, what are control strategies *for* on Twitter specifically. Users worked toward message dissemination bringing together like-minded users around a cause or shared interest. They defined groups within the boundary-less public space of Twitter. We call this assembly work drawing on how it is characterized in legal scholarship. In the absence of ways of explicitly configuring this, users instead resorted to manipulating the opaque processes of platform algorithms to bring about desired effects. This sometimes required that they work against platform configurations that prioritized engagement.

User-driven campaigns to 'deplatform' certain Twitter-using subcommunities which also surfaced in our dataset are trickier to reconcile with a notion of a right to assemble online. Here we see Twitter users not simply creating and protecting their own groups, but disrupting and undermining others. For example, a tweet thread started with the warning about pedophiles from Tumblr who are making their way to Twitter. The tweet identifies a long list of suspected accounts and defines coded terminology used by that community including "map" ("minor attracted person") and "nomap" ("non-offending minor attracted person"). The thread encouraged others to block these accounts and then use the "report" function on their tweets (in an attempt to engage content moderation processes). One tweet urges in reply that: "we have to keep reporting them. They should never be allowed to form a 'community' of any sort on Twitter. If we keep reporting, they'll probably disband." The following tweet found its way into our sample invoking the 'algorithm' as something they should strategize around despite its opaque workings:

> "I think child abuse fits into this category. So I've been reporting them as threatening violence or physical harm. Though I'm just not sure what's most effective when it comes to twitter's algorithm"

In these and other examples, users take on the role of platform arbiter when they perceive that Twitter has been negligent in this role. There are further examples of this in the scholarly literature. Stuart Geiger details the use of the Twitter API to build automated 'blocklists' – when one member blocks an account it is automatically blocked for all other members of the group [20]. This mass-blocking feature is one that Twitter has, to date, declined to implement. After users took the initiative, this tool proved to be an effective response to mob-based harassment.

In existing legal scholarship Inazu [30] calls these efforts at forming online social groups, "virtual assembly" referencing a lesser known clause from the US Constitution defending the "freedom of assembly." More specifically, he details four distinct "expressive acts" that enable this particular freedom. These include, "exclusion, embrace, expulsion, and establishment." We recognize Twitter

platform practices of user control as constituting these expressive acts. In acts of 'exclusion' Twitter users configure who will and will not be able to find and join their conversations, for example through coded language that helps them to avoid attention from critics or in building shared blocklists to disrupt mob-based harassment. In acts of 'embrace' Twitter users seek to reinforce relationships on the platform. Tweets we coded as "relationship management," show clearly that use of the platform is not simply about finding engaging content, or disseminating messages, but also building connections with other users. In acts of 'establishment' such as the campaigns of Korean boy band fans, they demonstrate user control efforts as part of the pursuit of shared goals.

4.6 Discussion

We can contextualize algorithmic control efforts on Twitter by noting a couple of transecting trends in social media. One is the way users are becoming concentrated on a small number of massively-scaled, social media platforms operated by the private sector. As Tufekci notes, such platforms are coming to serve as a *de facto* public sphere that, for global social movements, may be difficult to avoid as alternatives are absorbed through corporate acquisitions [46]. These platforms are becoming unavoidable for those seeking to disseminate an important message or pursue a social cause. A second trend is the dissolving distinction between online and offline groups. Platforms like Twitter are core spaces where social gathering takes place, not a simulacrum inferior to or with no bearing upon "real life." There are material differences in how an online public sphere and an offline one operate, but they do not exist as a hierarchy. In light of these two trends, arguments about private sector prerogative are weaker now than they perhaps were in the past.

On Twitter, users seek control over algorithmic processes to draw boundaries around the communities they are building, to protect them from abuse and harassment. They seek to overcome algorithmic errors, specifically failures to detect context and nuance. They resist, at times, the platforms relentless push toward engagement. Some users avoid being *known* by the algorithm, in a principled stance of resistance against the privacy intrusive background logging of their behavioral traces. These users obfuscate the coherent classification of their identities by a powerful automated system. These are some of the strategies that come to the fore when we take the perspective of users on control rather than that of the platform.

Our dataset surfaced the *social* processes animating social media. This was apparent in the creation of subgroups and communities around which boundaries were drawn through emergent shared language. From their Twitter accounts, these users build and tend to their relationships online, ones that algorithmic processes operating in the background may serendipitously facilitate, or conversely, may inhibit and damage. Twitter users try to direct the behavior of others to shape algorithmic processes in desired ways through admonishments and instructions that exert social pressure.

While control strategies that provide explicable and defensible social utility have been the focus of this paper, it is understandable that platforms like Twitter use algorithmic processes to manage challenges of scale and are concerned with preventing 'gaming.' The use of indirect and behavioral signals as inputs into these processes is a way to manage the presence of disruptive or resource-monopolizing actors, adversarial dynamics, and conflicts between users. It is typical for platforms to conceal details about the system's inner-workings to disarm those who seek to manipulate algorithmic processes for personal gain. However, a consequence of this is the shift toward regimes of technocratic governance online where users' experiences are shaped by the expertise and judgment of developers and corporate product teams. To lean too far in that direction is to undermine a broader distribution of control, including control granted to users.

By pursuing insight into algorithmic processes, users seek not just to satisfy their curiosity, but also to participate in determining the distribution of the platforms scarce resource: attention. This

does not seem to be facilitated readily by the Twitter platform. The marginality and uncertainty users express when they seek to exert control signals some of the hazards of the emerging public sphere role played by dominant online platforms. Major platforms concentrate considerable power over what has rapidly become an important public commons. The coordinated and algorithmically-aware efforts of users to manage attentional resources is an important new way freedom of assembly, among other forms of expression, is articulated in the present day.

5 CONCLUSION

Studying Twitter, admittedly as a convenient source for public social media data, allowed us to bring more voices into the debate about values in algorithmic systems. Focusing on control from the perspective of users broadened our scope of understanding beyond platform-centric considerations about the legitimacy of control and a focus on preventing 'gaming.'

Control is an important value to consider alongside fairness. This is not simply because giving users or other humans control over systems can make systems operate more effectively or efficiently (as where cognitive advantages are leveraged by pairing humans and machines through intelligence augmentation) but because it ties to fundamental dignitary concerns [31, 33]. On Twitter, algorithmic processes play an often quite visible role in content curation and, to a lesser extent, in content moderation. We identify both complaints about control and efforts to exert control by users as sometimes working against what the platform's business interests dictate (maximizing user engagement, preserving ad revenue). Upholding or protecting control we find reinforces other values and ideals that law scholars have defended such as "the right to a (willing) audience" [11] and freedom of association or assembly [30].

However, we also wish to recognize that savvy users manipulating curation algorithms or outwitting content moderation processes can also facilitate disturbing and disruptive phenomena on social media platforms such as mob-based harassment campaigns. And yet, similar forms of manipulation can also be used to fight back and protect users from such campaigns. We also recognize that simply giving those who are subject to algorithmic classification more control over these systems can be burdensome, a way of relinquishing responsibility for genuine problems that platforms ought to tackle themselves. Our goal in this paper was principally to broaden what we understand as the value of user control over algorithmic processes. We hope to spark deeper conversations and to push the needle on where the threshold between too much and too little control may lie.

6 ACKNOWLEDGEMENTS

We wish to thank the Algorithmic Fairness and Opacity Working Group (AFOG) at UC-Berkeley and its' members for stimulating conversations that led to this paper. We also wish to acknowledge a research gift from the Trust and Safety and Privacy and Security teams at Google Corporation that funded and facilitated AFOG.

REFERENCES

[1] Sara Ahmed. 2018. Refusal, Resignation and Complaint. Retrieved April 3, 2019 from https://feministkilljoys.com/2018/06/28/refusal-resignation-and-complaint/

[2] Jane Bambauer and Tal Zarsky. 2018. The Algorithm Game. *Notre Dame Law Review* 94, 1 (2018), 1–48. https://doi.org/10.1016/j.engfracmech.2007.08.004

[3] Solon Barocas and Andrew Selbst. 2016. Big Data's Disparate Impact. *California Law Review* 104 (2016).

[4] James R. Beniger. 1989. *The Control Revolution: Technological and Economic Origins of the Information Society.* Harvard University Press, Cambridge, MA.

[5] Reuben Binns. 2018. Fairness in Machine Learning : Lessons from Political Philosophy. In *Proceedings of Machine Learning Research, 2018 Conference on Fairness, Accountability, and Transparency.* 149–159.

[6] Sophie Bishop. 2018. Anxiety, panic and self-optimization: Inequalities and the YouTube algorithm. *Convergence* 24, 1 (2018), 69–84. https://doi.org/10.1177/1354856517736978

[7] Finn Brunton and Helen Nissenbaum. 2015. *Obfuscation: a user's guide for privacy and protest.* The MIT Press, Cambridge, MA.

[8] Taina Bucher. 2017. The algorithmic imaginary: exploring the ordinary affects of Facebook algorithms. *Information Communication and Society* 20, 1 (2017), 30–44. https://doi.org/10.1080/1369118X.2016.1154086

[9] Taina Bucher. 2018. Cleavage-Control: Stories of Algorithmic Culture and Power in the Case of the YouTube "Reply Girls". In *A Networked Self and Platforms, Stories, Connections.* Routledge, New York and London, 125–143.

[10] Alina Campan, Tobel Atnafu, Traian Marius Truta, and Joseph Nolan. 2019. Is Data Collection through Twitter Streaming API Useful for Academic Research? *Proceedings - 2018 IEEE International Conference on Big Data, Big Data 2018* (2019), 3638–3643. https://doi.org/10.1109/BigData.2018.8621898

[11] Jennifer A Chandler. 2007. A Right to Reach an Audience: An Approach to Intermediary Bias on the Internet. *Hofstra Law Review* 35 (2007), 1095–1138.

[12] Kathy Charmaz. 2006. *Constructing Grounded Theory: a practical guide through qualitative analysis.* Sage Publications, Thousand Oaks, CA.

[13] Alexandra Chouldechova. 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data* 5, 2 (2017), 1–6. arXiv:1703.00056 http://arxiv.org/abs/1703.00056

[14] Kelley Cotter. 2018. Playing the visibility game: How digital influencers and algorithms negotiate influence on Instagram. https://doi.org/10.1177/1461444818815684

[15] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Rich Zemel. 2011. Fairness Through Awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference.* Cambridge, MA, 214–226. https://doi.org/10.1145/2090236.2090255 arXiv:1104.3913

[16] Motahhare Eslami, Aimee Rickman, Kristen Vaccaro, Amirhossein Aleyasen, Andy Vuong, Karrie Karahalios, Kevin Hamilton, and Christian Sandvig. 2015. "I always assumed that I wasn't really that close to [her]". *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems - CHI '15* (2015), 153–162. https://doi.org/10.1145/2702123.2702556

[17] Motahhare Eslami, Kristen Vaccaro, Karrie Karahalios, and Kevin Hamilton. 2017. "Be careful; things can be worse than they appear": Understanding Biased Algorithms and Users' Behavior around Them in Rating Platforms. In *Proceedings of ICSMW.* https://www.aaai.org/ocs/index.php/ICWSM/ICWSM17/paper/viewPaper/15697

[18] Casey Fiesler and Nicholas Proferes. 2018. "Participant" Perceptions of Twitter Research Ethics. *Social Media and Society* 4, 1 (2018). https://doi.org/10.1177/2056305118763366

[19] Batya Friedman and Helen Nissenbaum. 1996. Bias in computer systems. *ACM Transactions on Information Systems* 14, 3 (jul 1996), 330–347. https://doi.org/10.1145/230538.230561

[20] R Stuart Geiger. 2016. Bot-based collective blocklists in Twitter: the counterpublic moderation of harassment in a networked public space. *Information Communication and Society* 19, 6 (2016), 787–803. https://doi.org/10.1080/1369118X.2016.1153700

[21] Tarleton Gillespie. 2017. Algorithmically recognizable: Santorum's Google problem, and Google's Santorum problem. *Information Communication and Society* 20, 1 (2017), 63–80. https://doi.org/10.1080/1369118X.2016.1199721

[22] Kevin D Haggerty and Richard V Ericson. 2002. The surveillant assemblage. *British Journal of Sociology* 51, 4 (2002), 605–622. https://doi.org/10.1080/00071310020015280

[23] Oliver Haimson and Anna Lauren Hoffmann. 2016. Constructing and enforcing "authentic" identity online: Facebook, real names, and non-normative identities. *First Monday* 21, 6 (2016).

[24] Moritz Hardt, Nimrod Megiddo, Christos Papadimitriou, and Mary Wootters. 2016. Strategic Classification. In *Proceedings of the 2016 ACM Conference on Innovations in Theoretical Computer Science.* 111–122. https://doi.org/10.1145/2840728.2840730 arXiv:1506.06980

[25] Moritz Hardt, Eric Price, and Nathan Srebro. 2016. Equality of Opportunity in Supervised Learning. In *Advances in Neural Information Processing Systems*, Vol. 29. 1–22. https://doi.org/10.1016/S0031-0182(03)00685-0 arXiv:1610.02413

[26] Eszter Hargittai. 2018. Potential Biases in Big Data: Omitted Voices on Social Media. *Social Science Computer Review* (2018), 1–15. https://doi.org/10.1177/0894439318788322

[27] Jeffrey Heer. 2019. Agency plus automation: Designing artificial intelligence into interactive systems. *Proceedings of the National Academy of Sciences* 116, 6 (2019), 1844–1850. https://doi.org/10.1073/pnas.1807184115

[28] Tad Hirsch, Kritzia Merced, Shrikanth Narayanan, Zac E Imel, and David C Atkins. 2017. Designing Contestability: Interaction Design, Machine Learning, and Mental Health.. In *DIS. Designing Interactive Systems (Conference)*, Vol. 2017. 95–99. https://doi.org/10.1145/3064663.3064703

[29] Eric Horvitz. 1999. Principles of mixed-initiative user interfaces. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems.* Pittsburgh, PA, 159–166. https://doi.org/10.1145/302979.303030

[30] John D Inazu. 2013. Virtual Assembly. *Cornell Law Review* 98, 5 (2013).

[31] Meg Leta Jones. 2017. The right to a human in the loop : Political constructions of computer automation and personhood. (2017). https://doi.org/10.1177/0306312717699716

[32] Matthew Joseph, Michael Kearns, Jamie Morgenstern, Seth Neel, and Aaron Roth. 2016. Rawlsian Fairness for Machine Learning. In *3rd Workshop on Fairness, Accountability, and Transparency Conference*. 1–26. arXiv:arXiv:1610.09559v1

[33] Margot E Kaminski. 2019. Binary Governance: Lessons from the GDPR's approach to algorithmic accountability. *Southern California Law Review* 92, 6 (2019).

[34] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2018. Inherent Trade-Offs in the Fair Determination of Risk Scores. In *Proceeding of the 2018 ACM International Conference on Measurement and Modeling of Computer Systems, SIGMETRICS*. Irvine, CA, 1–23. arXiv:1609.05807 http://arxiv.org/abs/1609.05807

[35] Min Kyung Lee and Su Baykal. 2017. Algorithmic Mediation in Group Decision: Fairness Perceptions of Algorithmically Mediated vs. Discussion-Based Social Division. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. Portland, OR, 1035–1048.

[36] Min Kyung Lee, Daniel Kusbit, Evan Metsky, and Laura Dabbish. 2015. Working with Machines: The Impact of Algorithmic and Data-Driven Management on Human Workers. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. Seoul, Republic of Korea, 1603–1612. https://doi.org/10.1145/2702123.2702548

[37] Bruno Lepri, Nuria Oliver, Emmanuel Letouze, Alex Pentland, and Patrick Vinck. 2017. Fair, transparent and accountable algorithmic decision-making processes: The premise, the proposed solutions, and the open challenges. *Philosophy & Technology* 31, 4 (2017), 611–627.

[38] Alice E. Marwick and Danah Boyd. 2011. I tweet honestly, I tweet passionately: Twitter users, context collapse, and the imagined audience. *New Media and Society* 13, 1 (2011), 114–133. https://doi.org/10.1177/1461444810365313 arXiv:arXiv:1011.1669v3

[39] Deirdre K Mulligan and Daniel S Griffin. 2018. Rescripting Search to Respect the Right to Truth. *Georgetown Law Technology Review* 557 (2018), 557–584.

[40] Gina Neff and Peter Nagy. 2016. Automation, Algorithms, and Politics| Talking to Bots: Symbiotic Agency and the Case of Tay. *International Journal of Communication* 10, 0 (2016), 17.

[41] Kimberly A. Neuendorf. 2002. *The Content Analysis Guidebook*. Sage Publications, Thousand Oaks, CA.

[42] Helen Nissenbaum. 2001. How Computer Systems Embody Values. *Computer* (2001), 118–120.

[43] Andrew D. Selbst and Solon Barocas. 2018. The Intuitive Appeal of Explainable Machines. *Fordham Law Review* 1085 (2018). https://doi.org/10.2139/ssrn.3126971

[44] Tao Stein, Erdong Chen, and Karan Mangla. 2011. Facebook immune system. *Proceedings of the 4th Workshop on Social Network Systems - SNS '11* m (2011), 1–8. https://doi.org/10.1145/1989656.1989664

[45] Zeynep Tufekci. 2014. Big Questions for Social Media Big Data: Representativeness, Validity and Other Methodological Pitfalls. *Proceedings of the 8th International AAAI Conference on Weblogs and Social Media* (2014), 505–514. https://doi.org/10.1016/j.ecoleng.2016.05.077 arXiv:1403.7400

[46] Zeynep Tufekci. 2017. *Twitter and Tear Gas: The Power and Fragility of Networked Protest*. Yale University Press, New Haven and London.

[47] Emily van der Nagel. 2018. 'Networks that work too well': intervening in algorithmic connections. *Media International Australia* 168, 1 (2018), 81–92. https://doi.org/10.1177/1329878X18783002

[48] Jeffrey Warshaw, Nina Taft, and Allison Woodruff. 2016. Intuitions, Analytics, and Killing Ants: Inference Literacy of High School-educated Adults in the US. In *Proceedings of the Twelfth Symposium on Usable Privacy and Security*. Denver, CO.

[49] Michele Willson. 2017. Algorithms (and the) everyday. *Information Communication and Society* 20, 1 (2017), 137–150. https://doi.org/10.1080/1369118X.2016.1200645

[50] Kevin Witzenberger. 2018. The Hyperdodge: How Users Resist Algorithmic Objects in Everyday Life. *Media Theory* 2, 2 (2018), 29–51. http://journalcontent.mediatheoryjournal.org/index.php/mt/article/view/56/46

[51] Allison Woodruff, Sarah E. Fox, Steven Rousso-Schindler, and Jeffrey Warshaw. 2018. A Qualitative Exploration of Perceptions of Algorithmic Fairness. 1–14. https://doi.org/10.1145/3173574.3174230

A CODEBOOK

(F: frequency of the code in our dataset sample, K: kappa coefficient score)

META-CATEGORY: EXPLANATION

Theme	Description
seeking explanation or offering explanation (F: 35%, K: 0.504)	Puzzling over the algorithm, seeking explanation, or explaining the algorithm to others
sense-making (F: 12.5%, K: 0.804)	Observing instances where the algorithm reveals its logic, often funny juxtapositions, the algorithm is 'amusing,' delighting in discovery of how the algorithm works, but agnostic about whether the algorithm is malfunctioning

META-CATEGORY: COMPLAINTS

Theme	Description
algorithm is in error or is awry (F: 13.8%, K: 0.732)	Content that is insufficiently personalized, Twitter algorithm seems to say things about me that I find unflattering, algorithm did something that is laughably wrong, algorithm is underwhelming, explicit or implicit comparisons to human judgment where algorithm is deemed inferior
accusing of some form of bias (F: 5.77% K: 0.884)	Beyond complaints about personal attention to imply or express that a group or category of users is systematically being denied attention by the platform, i.e. bias against conservatives, bias against women, bias in burying content from users in certain geographic regions
complaining about what is or isn't seen (F: 26.86% K: 0.559)	Intrusively exposed to or pestered by unwelcome or repetitive content, desired content is hidden, user's own tweets seem hidden or are getting limited attention, accusations of censorship
lamenting lack of control (F: 12.66% K: 0.685)	Complaining about being controlled by the algorithm, censorship of speech, feeling compelled to change behavior in undesired ways to appease algorithm

META-CATEGORY: RELATIONSHIP MANAGEMENT

Theme	Description
explaining self (F: 35.44% K: 0.826)	Proactive relationship management, explaining how one came to view another's content or became a follower, defending against potential accusations of creepiness
shifting blame from person to algorithm (F: 13.08% K: 0.783)	Moments where assumptions about human intent (someone is ignoring me) are shifted to the algorithm (not seeing tweets, because the algorithm excludes them from the timeline)
commiserating about the algorithm (F: 4.64% K: 0.873)	Responding to someone's complaint, trying to make them feel better. Using the algorithm to establish common ground, form a connection with another user

META-CATEGORY: EXERTING CONTROL OVER THE ALGORITHM

Theme	Description
exerting control, gaming (F: 17.02% K: 0.685)	Attempts to exert control over the algorithm, attempting to manipulate attention (or discussing how to do that), asking other users to assist in overcoming algorithm, 'correcting' the practices of other users with regards to Twitter algorithms

Received April 2019; revised June 2019; accepted August 2019